Fundamental Limits and Practical Codes for Collaborative Estimation

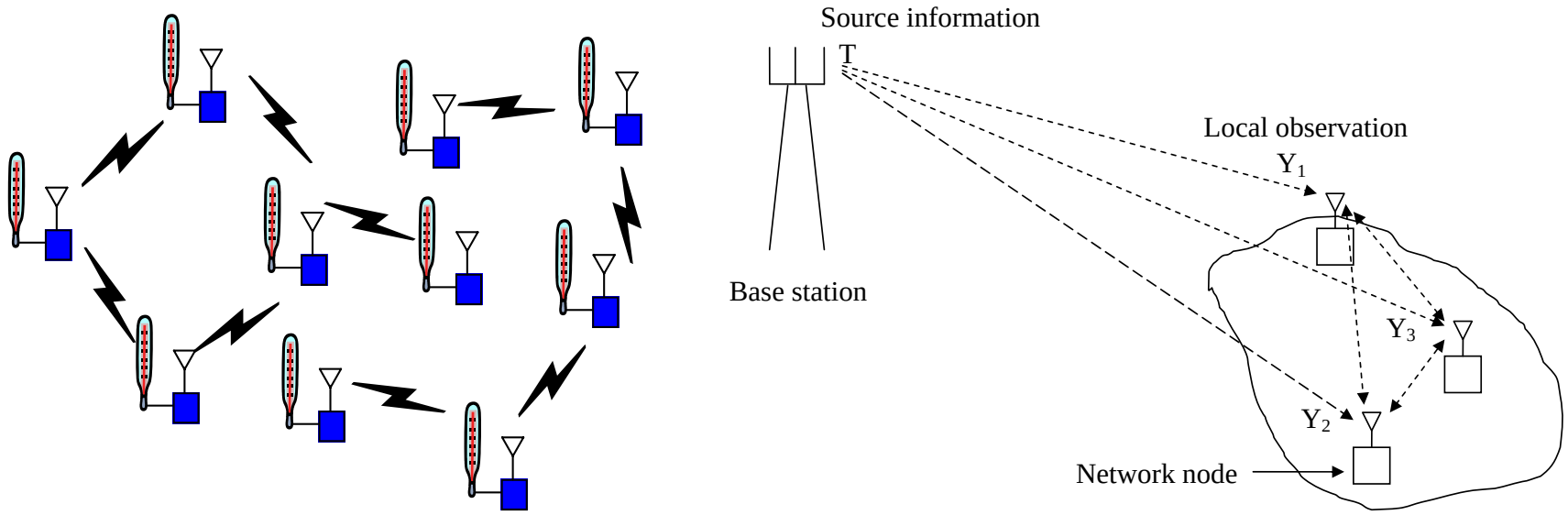
John MacLaren Walsh, Sivagnanasundaram Ramanan
Department of Electrical and Computer Engineering
Drexel University
Philadelphia, PA
jwalsh@ece.drexel.edu

Thanks to NSF CCF-0728496 & CCF-1016588.

Outline

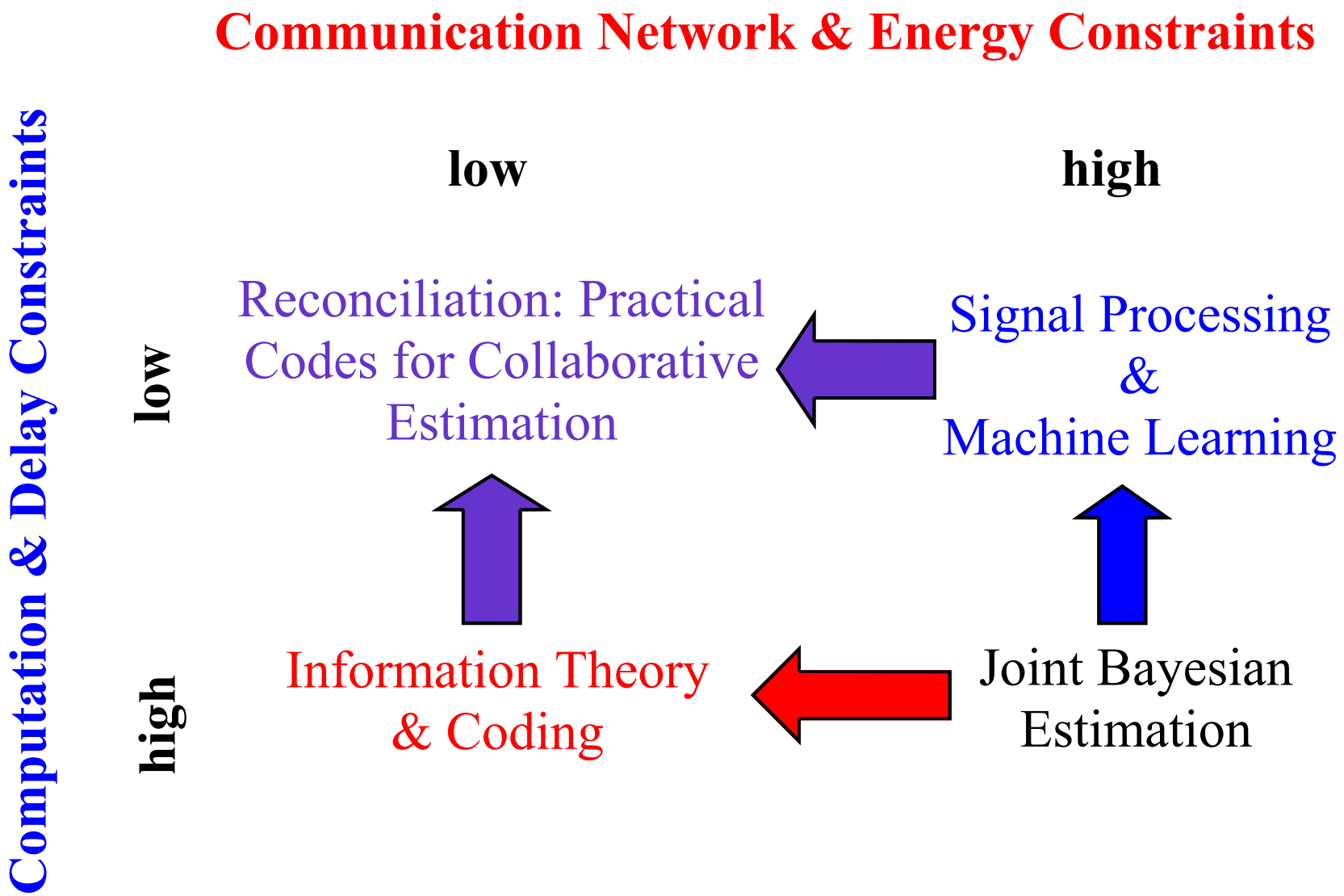
- Collaborative Estimation Problem
  - model set up
  - key constraints (complexity, communication), perspectives
- Dealing with Complexity via Approximate Inference
  - Information Geometry determines performance complexity tradeoff
- Dealing with Communication Limitations via Multi-terminal Rate Distortion Theory
  - Entropy Geometry determines performance communication tradeoff
- How to Reconcile to Deal with both Constraints
  - Couple Approximate Bayesian Joint Inference with BP Decoders
- Proof of Concept

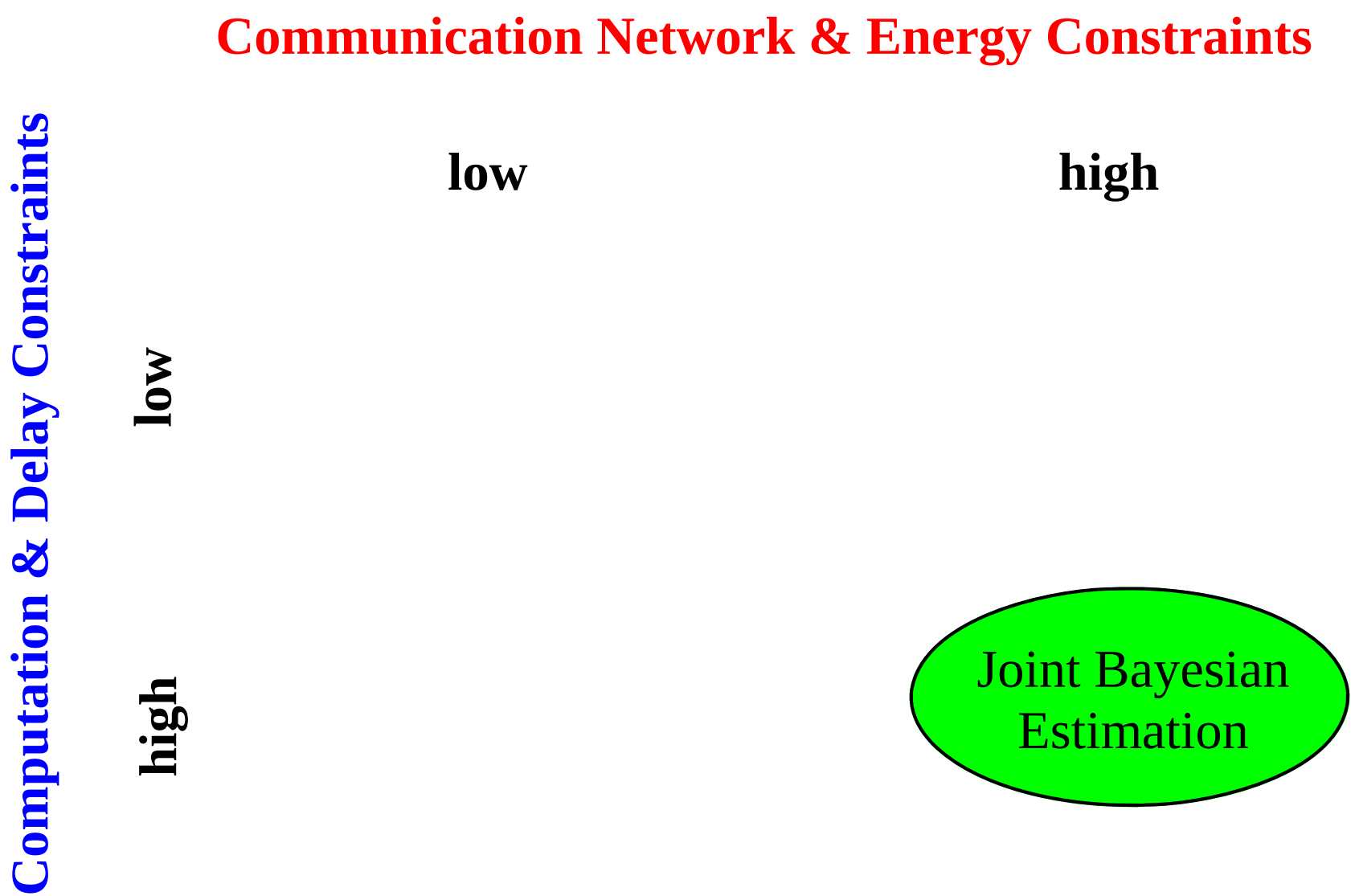
What is collaborative estimation?



- M nodes. Node m w/ local observations $\mathbf{R}_m$.
- Collection of random parameters $\boldsymbol{T}$ jointly distrib. w/ $\{\mathbf{R}_m\}$
- Node m wants to estimate $\boldsymbol{T}$ with $\hat{T}_m$ to minimize a local Bayesian cost function, i.e. $d_m(\hat{T}_m, \boldsymbol{T})$ given avail. info.
- nodes share information over a network to help form their estimates

What are the major research issues/perspectives?

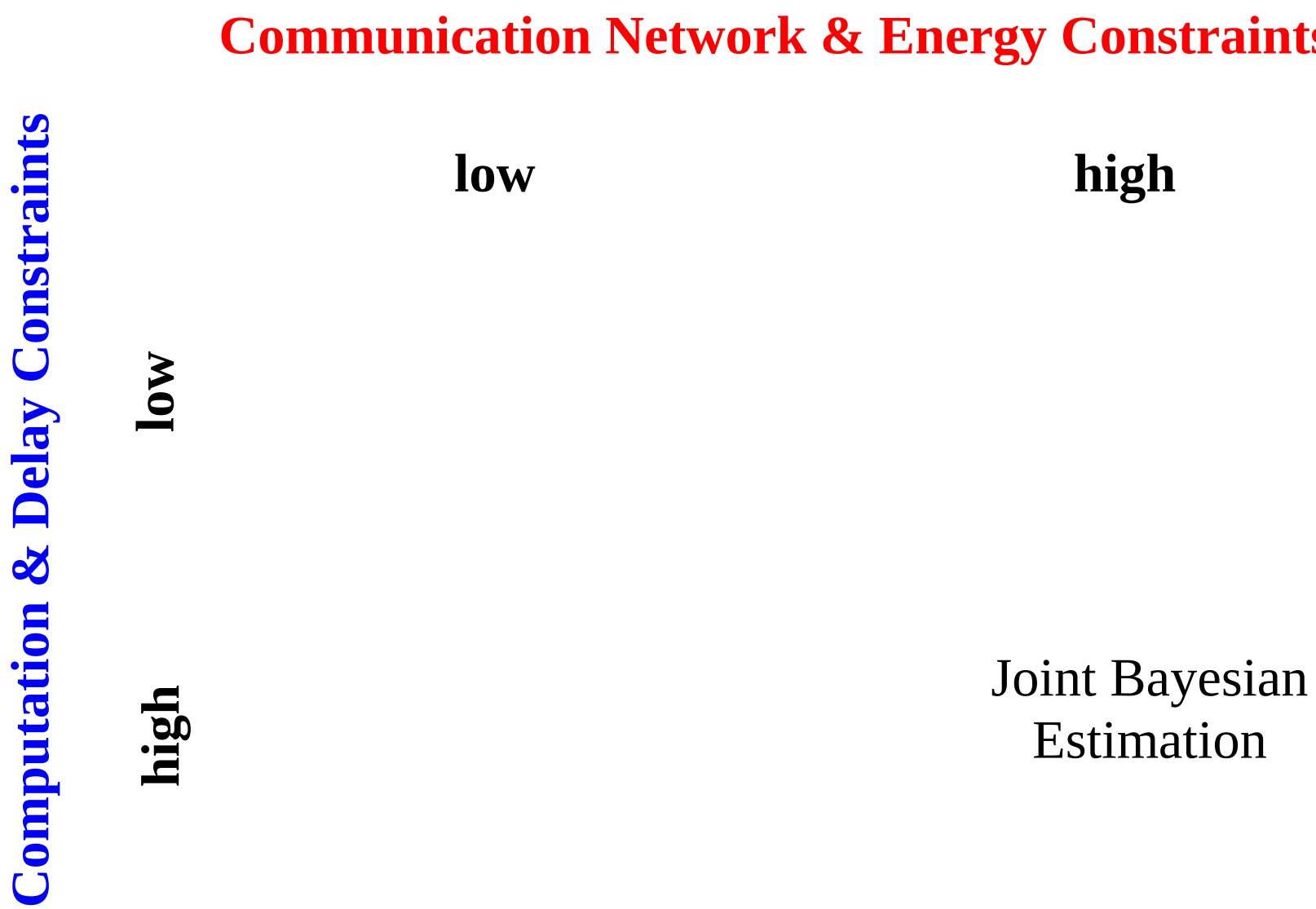



Communication Network & Energy Constraints
low
high
Computation & Delay Constraints
low
high
Joint Bayesian Estimation

Joint Bayesian Estimation

- *Without the constraints, the problem is trivial once the model has been selected.*
- each node broadcasts its observations $\mathbf{r}_m$ to all of the other nodes
- given $\mathbf{r} := [\mathbf{r}_m | m \in [M]]$ each node forms the posterior distribution $p_{\mathbf{T}|\mathbf{R}}(\mathbf{T}|\mathbf{r})$.
- each node chooses its estimate $\hat{\mathbf{T}}_m$ as the estimate minimizing its own Bayesian risk function

$$\boxed{\hat{\mathbf{T}}_m \in \arg\min_{\hat{\mathbf{T}}_m} \int d_m(\hat{\mathbf{T}}_m, \mathbf{T}) p_{\mathbf{T}|\mathbf{R}}(\mathbf{T}|\mathbf{r}) \mathrm{d}\mathbf{T}}$$


Communication Network & Energy Constraints
low
high
Computation & Delay Constraints
low
high
Joint Bayesian Estimation

Communication Network & Energy Constraints

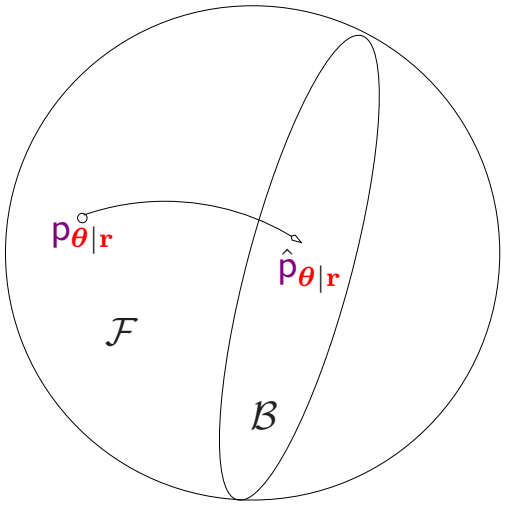
Computation & Delay Constraints

	low	**high**
low		Signal Processing & Machine Learning
high		Joint Bayesian Estimation

The Signal Processing/Machine Learning Perspective

PROBLEM 1: $\arg\min_{\hat{\mathbf{T}}_m} \int d_m(\hat{\mathbf{T}}_m, \mathbf{T}) p_{\mathbf{T},\mathbf{R}}(\mathbf{T}, \mathbf{r}) \mathrm{d}\mathbf{T}$ IS HARD!

- one important major difficulty: the integration over the posterior distribution is usually difficult computationally and analytically, as can be the minimization of the local risk.
  - enter approximate Bayesian inference: "best"-approximate the posterior distribution within a tractable family of distributions
    - Gibbs Sampling
    - Variational Bayes
    - Expectation/Belief Propagation



- Complexity handled in 2 respects
  - Approximating Family selected so that risk calc. & min. is easy
  - a factoring of $p(\boldsymbol{T}|\boldsymbol{r}) = \prod_a f_a(\boldsymbol{T}_a, \mathbf{r}_a)$ is exploited to individually fit factors of an approx. distr. $\hat{p}(\boldsymbol{T}|\boldsymbol{r}) = \prod_a g_a(\boldsymbol{T}_a)$ (yields a "message passing" interpretation)

How can this be used to simplify the Risk Minimization/Calculation?

$$\arg\min_{\hat{\mathbf{T}}_m} \int d_m(\hat{\mathbf{T}}_m, \mathbf{T}) \hat{p}_{\mathbf{T}|\mathbf{R}}(\mathbf{T}|\mathbf{r}) \mathrm{d}\mathbf{T}$$

- if $d_m(\mathbf{T}, \hat{\mathbf{T}}_m) = d_m(\mathbf{T}_m, \hat{\mathbf{T}}_m)$ can select a factoring & approximating family to get marginals for $\mathbf{T}_m$.
- If the risk is a sum of terms of this form, can again simply find the best marginal approx., i.e. $\hat{p}(\mathbf{T}|\mathbf{r}) = \prod_m \hat{p}(\mathbf{T}_M|\mathbf{r})$
- More broadly, if there are parts of the posterior which yield risk computation difficult, they can be approximated with exponential families in which it is simple (e.g. Gaussians). [1, 2]
- The message passing nature of the algorithm describes one way to handle decentralization of the data (group it with factor nodes). [3, 4, 5, 6, 7]

What is the major underlying fundamental (math) problem here?

- The selection of the factoring

$$p(\boldsymbol{T}|\boldsymbol{r}) = \prod_a f_a(\boldsymbol{T}_a, \mathbf{r}_a) \quad \& \text{ the approximating family } \mathcal{B}$$

  determines *both*:

  – the convergence & the **complexity** of the variational inference, as well as

  – the **performance** of the estimates (i.e. the error in the risk calculation)

- Further, there is a tension:

  – A bigger approximating family allows for equal or better *performance*, but comes at the cost of additional *complexity* in fitting the approximate distribution and calculating the risk.

  – Additionally, a bigger approximating family requires more parameters, and hence requires bigger messages, hence more communication.

- *Characterize this performance vs complexity tradeoff and approximating families which attain it.* (Dictated by interplay between parameterizations of subfamilies of distributions and estimate errors, i.e. information geometry [8, 9].)

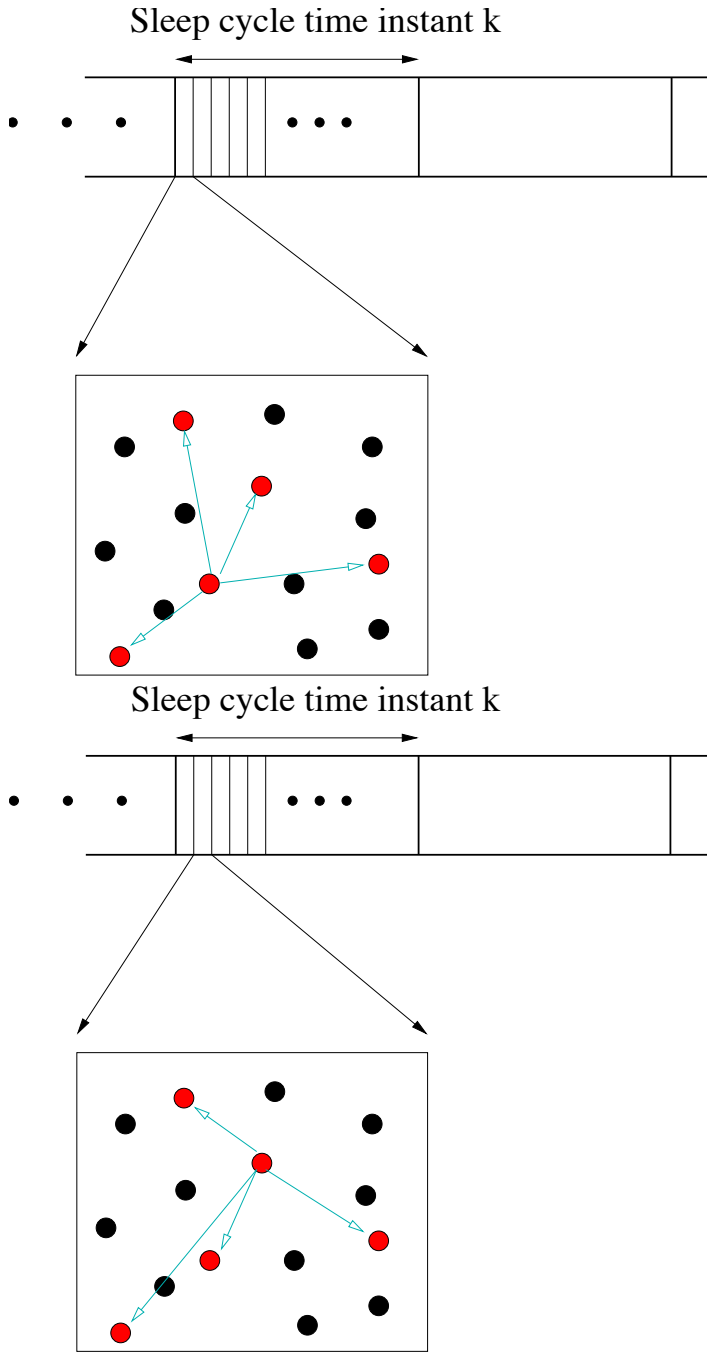
Let's See this in an Example

- N network nodes placed IID at positions $X_i \sim p_X$
- Maintain low power consumption – random duty cycling $\mathcal{A}(k)$
- Need to organize a wireless network between them (neighbors)
- Must estimate $(N-1)N$ channel SNRs/gains between them

$$h_{ij} \approx\propto \|X_i - X_j\|_2^{-\alpha} \qquad (1)$$

- Gives a (marginal) prior p_H that is intractable, but potentially useful
- Standard channel training, $w_{ijk} \sim \mathcal{N}(0,1)$

$$r_{ijk} = h_{ij} s_{ik} + w_{ijk} \qquad (2)$$



Let's See this in an Example - cont'd

- Posterior factors as

$$p(\mathbf{H}|\mathbf{r}) \propto p(\mathbf{H}) \prod_{ijk} p(r_{ijk}|h_{ij}) \qquad (3)$$

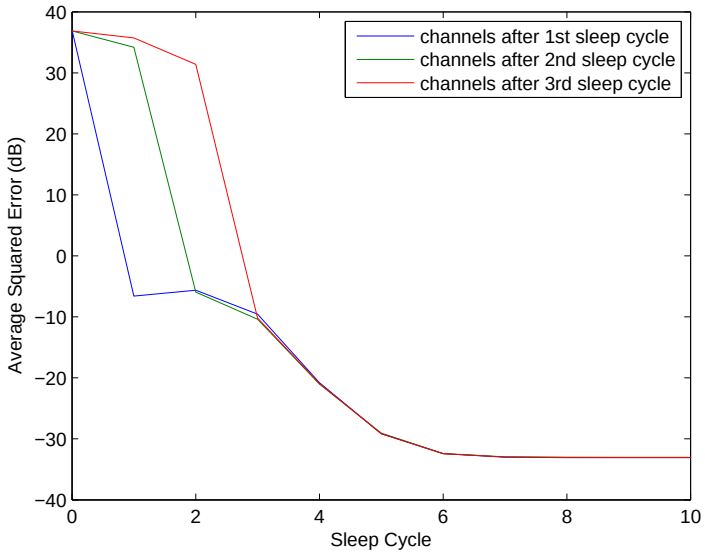
- Match with an approximate posterior

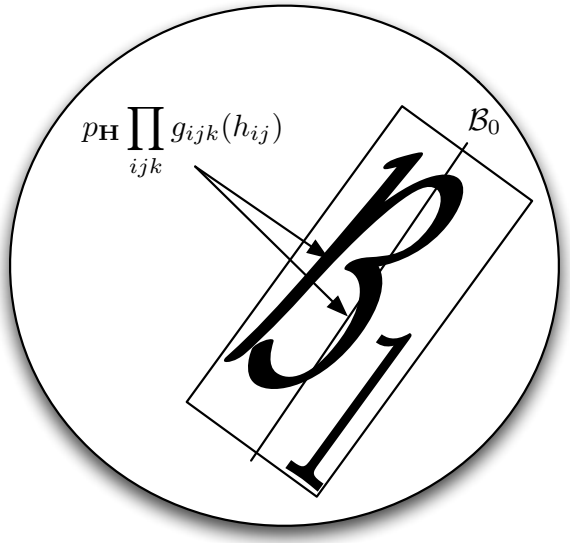
$$\hat{p}(\mathbf{H}|\mathbf{r}) \propto g_0(\mathbf{H}) \prod_{ijk} g_{ijk}(h_{ij}) \qquad (4)$$

- Expectation propagation refines approx. factors according to

$$g_a = \arg \min_{g_a \in \mathcal{B}_a} D\left(f_a \prod_{c \neq a} g_c \,\|\hat{p} \right) \qquad (5)$$

- [1, 2]

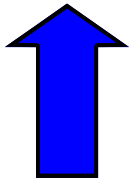


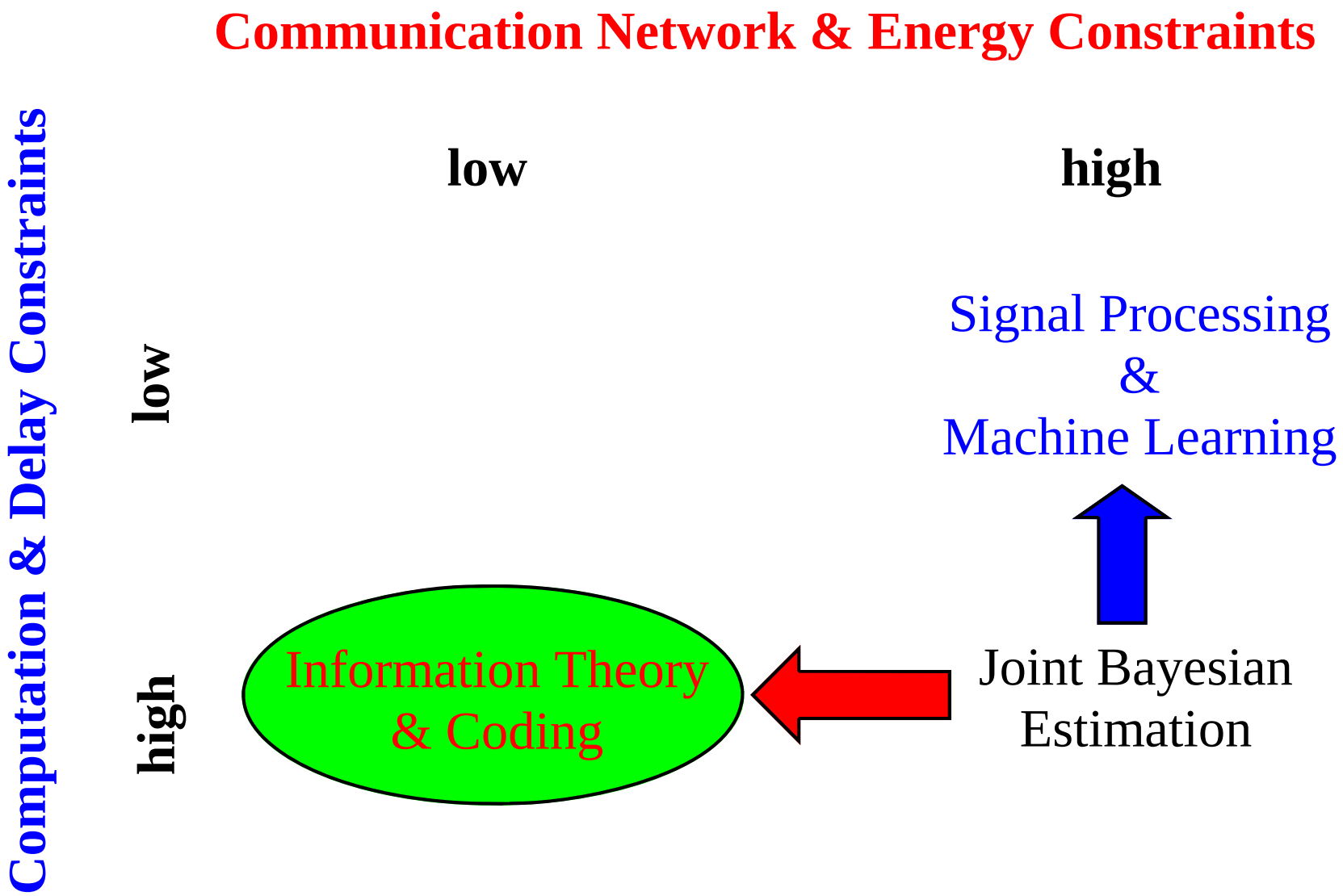


Communication Network & Energy Constraints

Computation & Delay Constraints

	low	**high**
low		Signal Processing & Machine Learning
high		Joint Bayesian Estimation



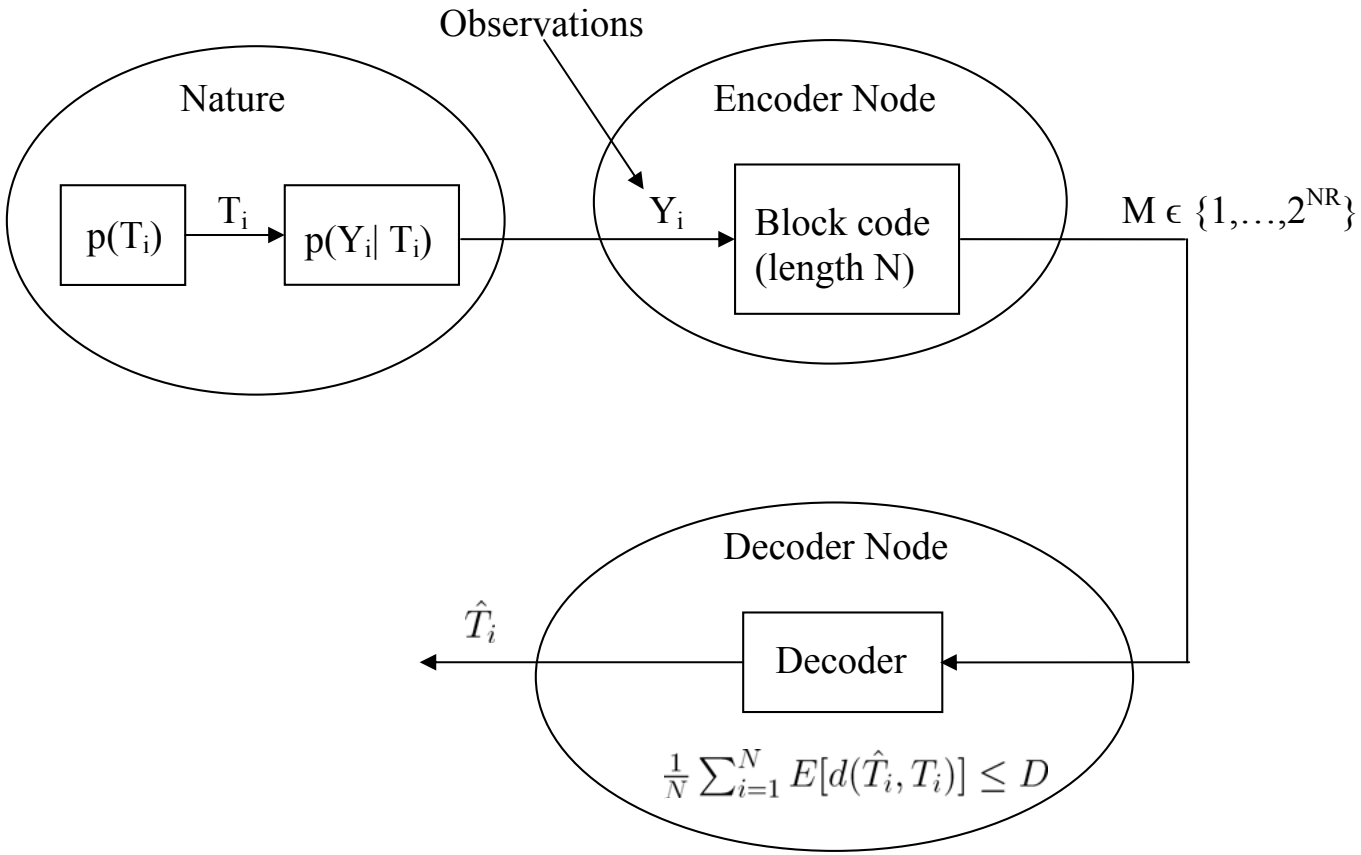

Communication Network & Energy Constraints
Computation & Delay Constraints
low
high
low
high
Signal Processing
&
Machine Learning
Information Theory
& Coding
Joint Bayesian
Estimation

The Information Theory/Coding Perspective

PROBLEM 2: only $\mathbf{r}_m$ is originally available to node m,
any communication is over finite rate links

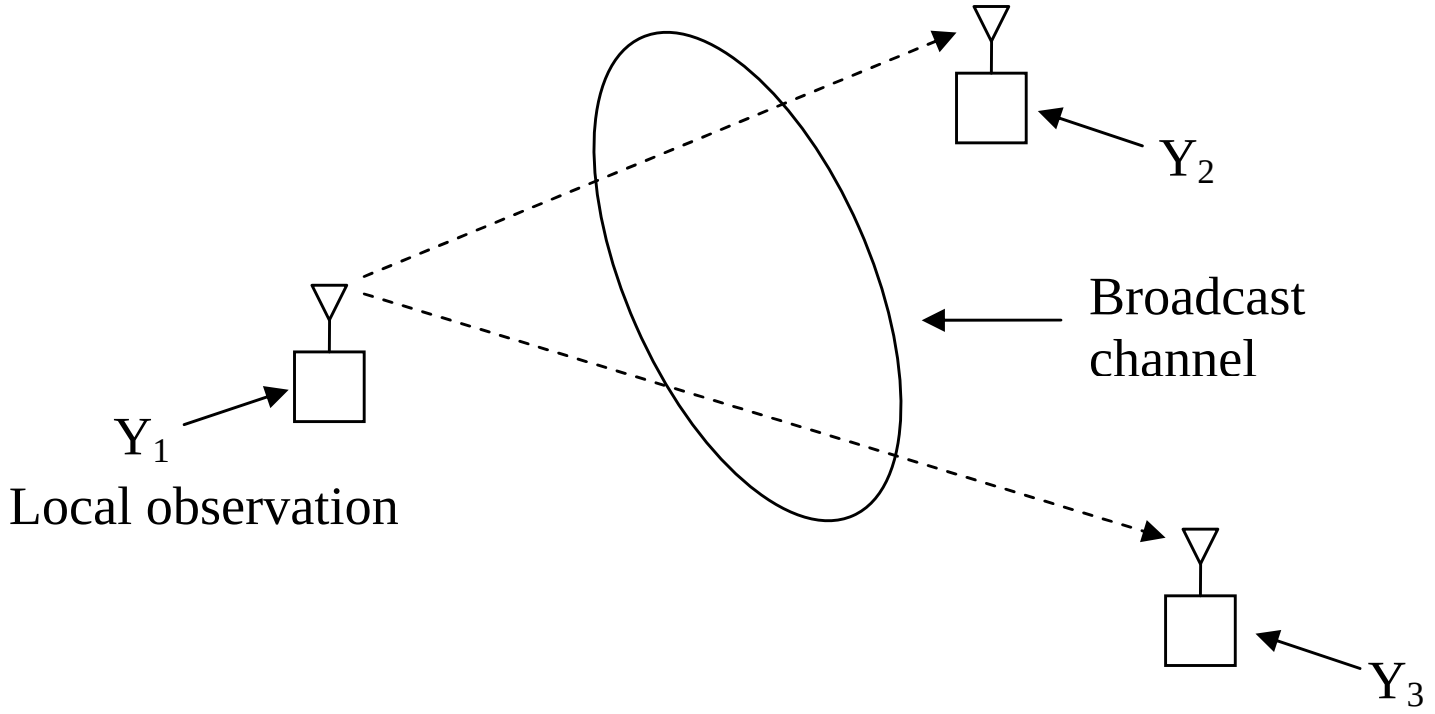
- another major important difficulty: the nodes must send their messages over a rate limited wired or wireless communication network. The information exchanged can not exceed the capabilities of this network, and the level of use may lead to costs (e.g. energy or $).
- How should the communications be organized to allow for the best estimate performance when adapted to different communications networks? (I.e. what is the code structure?)
- What is the *best* estimate performance we can have subject to these constraints?

Relationship Between Remote Bayesian Estimation and Lossy Source Coding



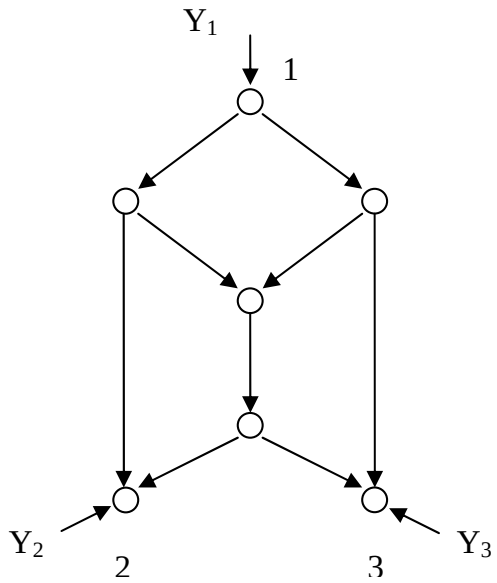
- $\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[d(\hat{T}_i, T_i)] < D$ plays the role of an average Bayesian cost. Dobrushin & Tsybakov '62 [10] showed minimum rate necessary to attain $< D$ is $\min I(U;Y)$ over $U \leftrightarrow Y \leftrightarrow T$.
- Just like rate distortion function but with T_i instead of Y_i in distortion, and Markov requirement.

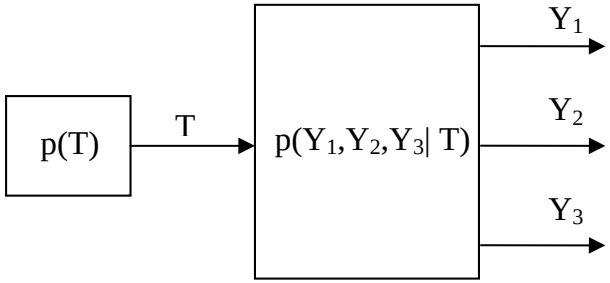
What should the source code architecture be under SC separation?



- Scalar Gaussian broadcast channel is degraded:
  - everything that receiver w/ $\downarrow$ SNR gets, the receiver with $\uparrow$ SNR gets
  - receiver w/ $\uparrow$ SNR can get extra info
- Source code construction should reflect this:
  - If source code sends only individual messages $S_{1\to 2}, S_{1\to 3}$ the ability of receiver w/ $\uparrow$ SNR to hear everything sent to the receiver w/ $\downarrow$ SNR is *wasted*
  - $\Rightarrow$ should use *multicast* messages! $S_{1\to\{2,3\}}, S_{1\to 2}, S_{1\to 2}$.

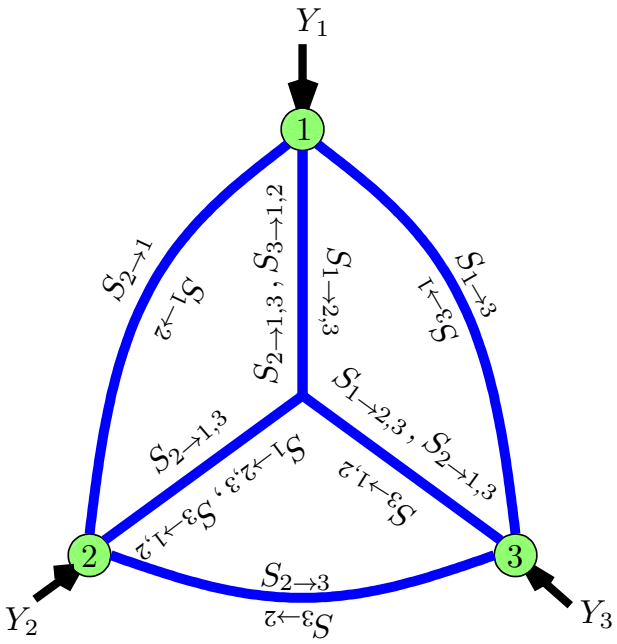
What should the source code architecture be?

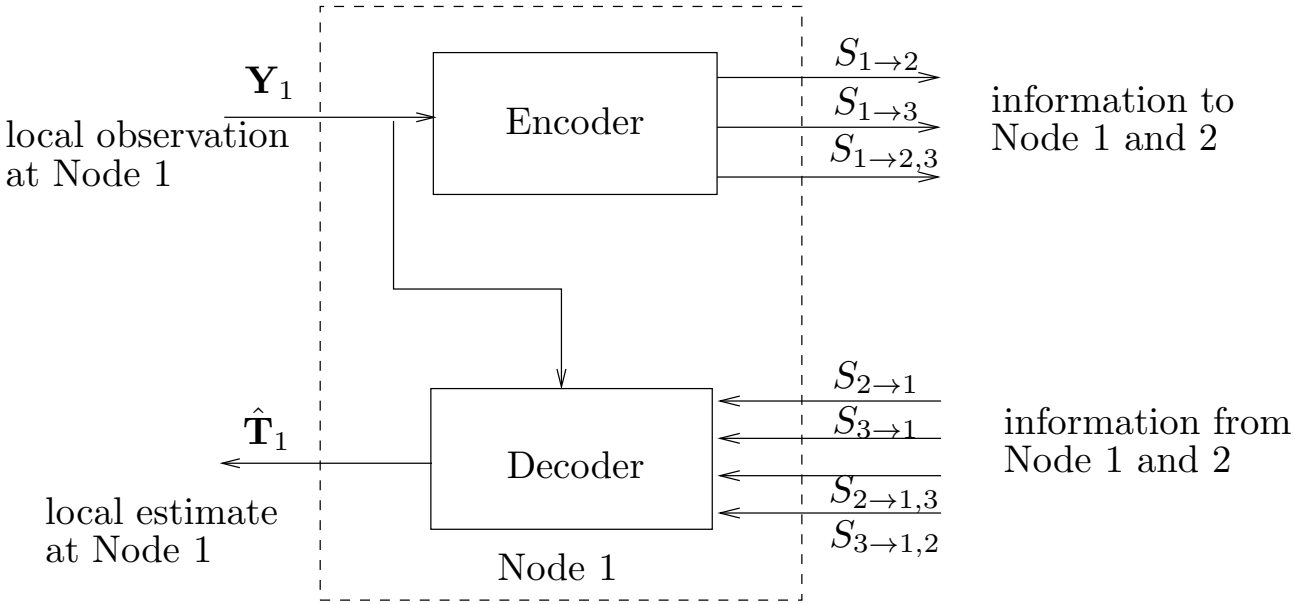




- Network coding insight: limitation for $R_{1\to\{2,3\}}$ is 2, higher than maximum equal $R_{1\to 2}, R_{1\to 3} = \frac{3}{2}$.
- Again implies that (even separated) source coding construction should allow for *multicast* rates.

What should the source code architecture be?





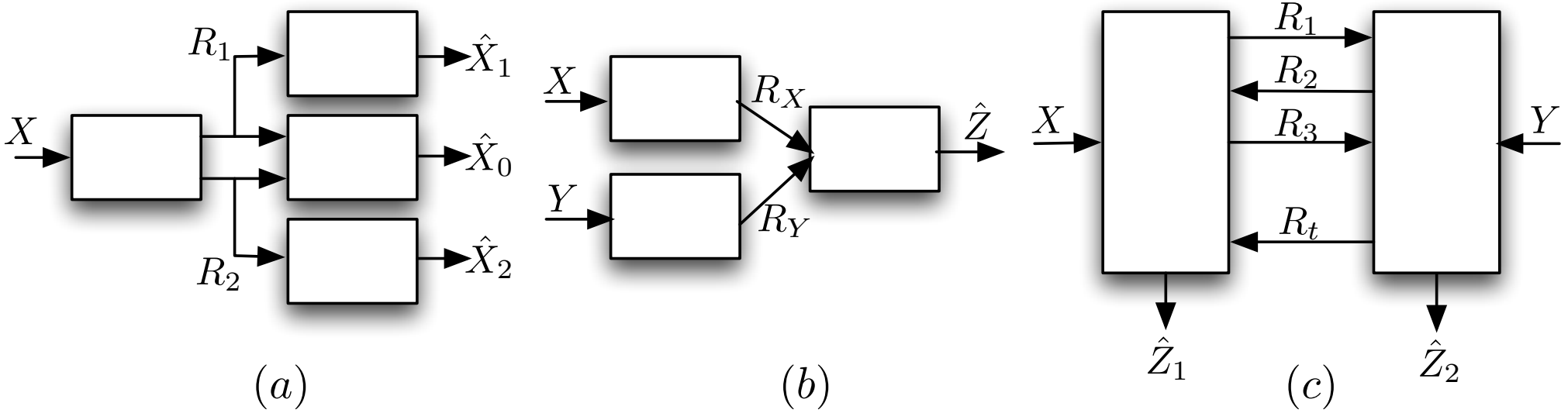
- Each node sends a (possibly different nonempty) message to each subset of other nodes. $S^k_{j\to\mathcal{A}}$, $\mathcal{A} \subseteq [M] \setminus j$. Multiple rounds k of such messages may be sent, with messages depending on messages received in previous rounds.
- Every node collects all of the received messages together with its local observations and forms an estimate which minimizes its local Bayesian cost $\mathbb{E}\left[d_m(T, \hat{T}_m)\right]$.

What performance do the best such codes have?

- Rate distortion region $\mathcal{R}$ of achievable rate vector $\mathbf{r} := [R_{j\to\mathcal{A}} | j \in [M],\ \mathcal{A} \subseteq [M] \setminus j]$ and estimation error (cost) vector $\mathbf{d} := [D_j | j \in [M]]$ pairs characterizes the best such codes.
- Capacity region $\mathcal{C}$ of a network is described in terms of all achievable $\mathbf{r}$.
- Estimation performances attainable are those $\mathbf{d}$ associated with a $\mathbf{r}$ through $(\mathbf{r}, \mathbf{d}) \in \mathcal{R}$ with $\mathbf{r}$ in $\mathcal{C}$.
- Hence, inner and outer bounds for the rate distortion region $\mathcal{R}$ for this problem are of interest.

Rate Distortion Region

this problem is a hybrid btw. 3 classic "incompletely solved" IT problem classes...[11][12][13] [14, 15]



- (a) - Multiple descriptions, (b) - CEO, (c) - interactive rate distortion
- "Incompletely solved" $=$ *single letter characterizations* are generally not available
- *If we don't ask for single letter*, can be expressed as

$$NR_{i,j}^k \geq H(S_{i\to\mathcal{A}}^k), \quad S_{i\to\mathcal{A}}^k \leftrightarrow Y_i^N \mathbf{S}_{\to i}^{k-1} \leftrightarrow \text{everything else} \tag{6}$$

- with $\hat{T}_i^N \leftrightarrow Y_i^N S_{\to i}^K \leftrightarrow$ everything else such that $\mathbb{E}[d_i(T^N, \hat{T}_i^N)] < D_i$

What is the major underlying fundamental (math) problem here?

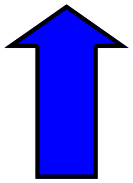
- Major issue with these regions: while analytically elegant, it can be quite difficult to determine whether or not a given numerical $\mathbf{r}, \mathbf{d}$ pair lies within them. (e.g. plot them or optimize over them)
- They involve inequalities among *rates* and (weighted) sums of *Shannon entropies* of subsets of random variables, including *auxiliary variables* (distribution not determined other than to obey certain distortion constraints & Markov conditions).
- All rate regions in multiterminal information theory are expressible in this way.
- Hence all rate regions are expressed in terms of linear projections of $\bar{\Gamma}^*_N(\mathcal{C})$.
- Problem is, we don't know the boundaries of $\bar{\Gamma}^*_4$, let alone $\bar{\Gamma}^*_N$ or $\bar{\Gamma}^*_N(\mathcal{C})$. [16, 17]
- Hence, underlying problem here is of determining *entropy geometry*.

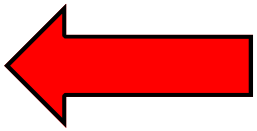
How might these perspectives be reconciled?

Communication Network & Energy Constraints

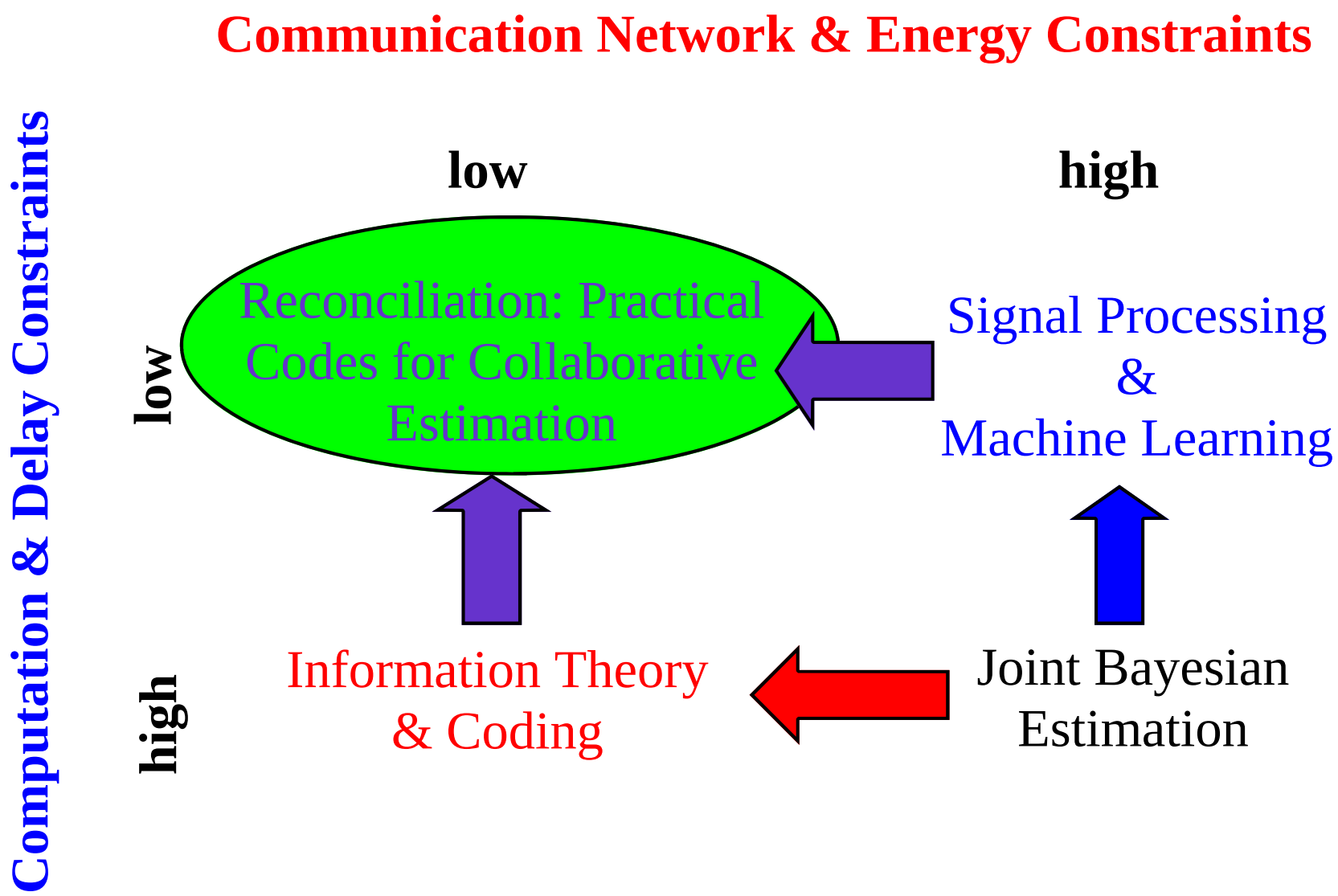
Computation & Delay Constraints

	low	**high**
low		Signal Processing & Machine Learning
high	Information Theory & Coding	Joint Bayesian Estimation





How might these perspectives be reconciled?

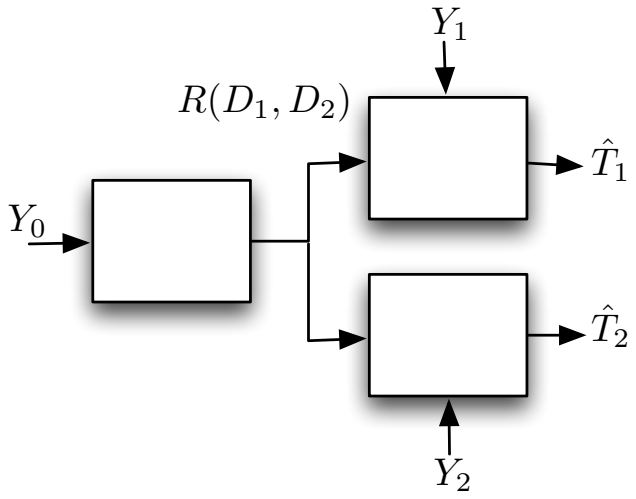


How might these perspectives be reconciled?

- Sparse graph coding constructions and modifications of BP decoders have been adapted to some multiterminal coding problems (Wyner-Ziv, Slepian-Wolf)
- How can they be adapted and generalized to this one?
- What do the information theoretic bound evaluate to in important pragmatic estimation problems for wireless networks, such as for channel estimation?
- Belief/expectation propagation can help not only with designing the decoders, but also determining which information to compress in order to make risk minimization tractable after decoding.

How might we do this? – Information Theory Part

- simplify to broadcast messages
- send messages one at a time
- essential problem here: broadcasting a message to nodes with differing levels of side information
- Direct *degraded* case $Y_0 \leftrightarrow Y_1 \leftrightarrow Y_2$ established by *Source Coding when Side Information May be Absent*, C. Heegard and T. Berger, Trans. I. T., 1984.
- Indirect degraded case $T \leftrightarrow Y_0 \leftrightarrow Y_1 \leftrightarrow Y_2$ follows from this with not much work.



Heegard Berger Problem

$$R \geq I(Y_0; U_1|Y_1) + I(Y_0; U_2|U_1 Y_2) \quad (7)$$

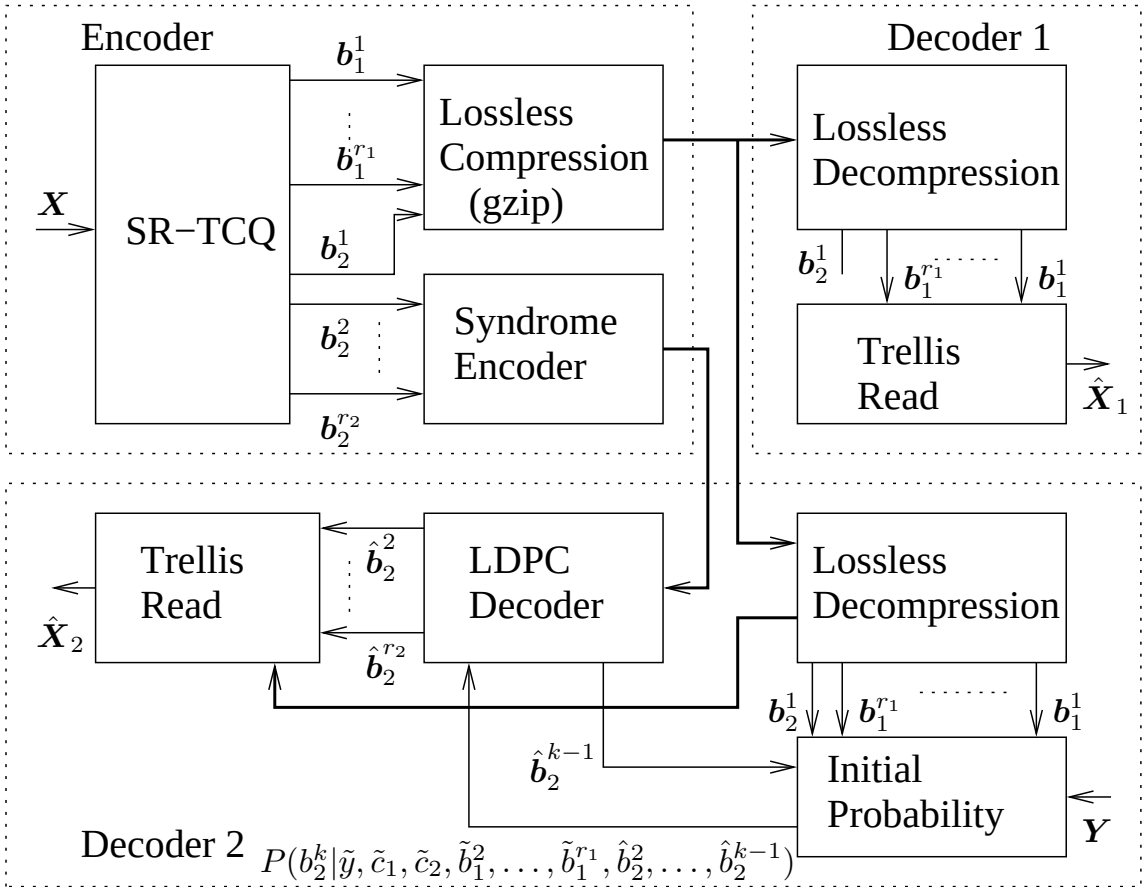
$$\mathbb{E}[d_1(T; g_1(Y_1, U_1))] \leq D_1 \quad (8)$$

$$\mathbb{E}[d_2(T; g_2(Y_2, U_1, U_1))] \leq D_2 \quad (9)$$

$$U_1, U_2 \leftrightarrow Y_0 \leftrightarrow Y_1, Y_2 \quad (10)$$

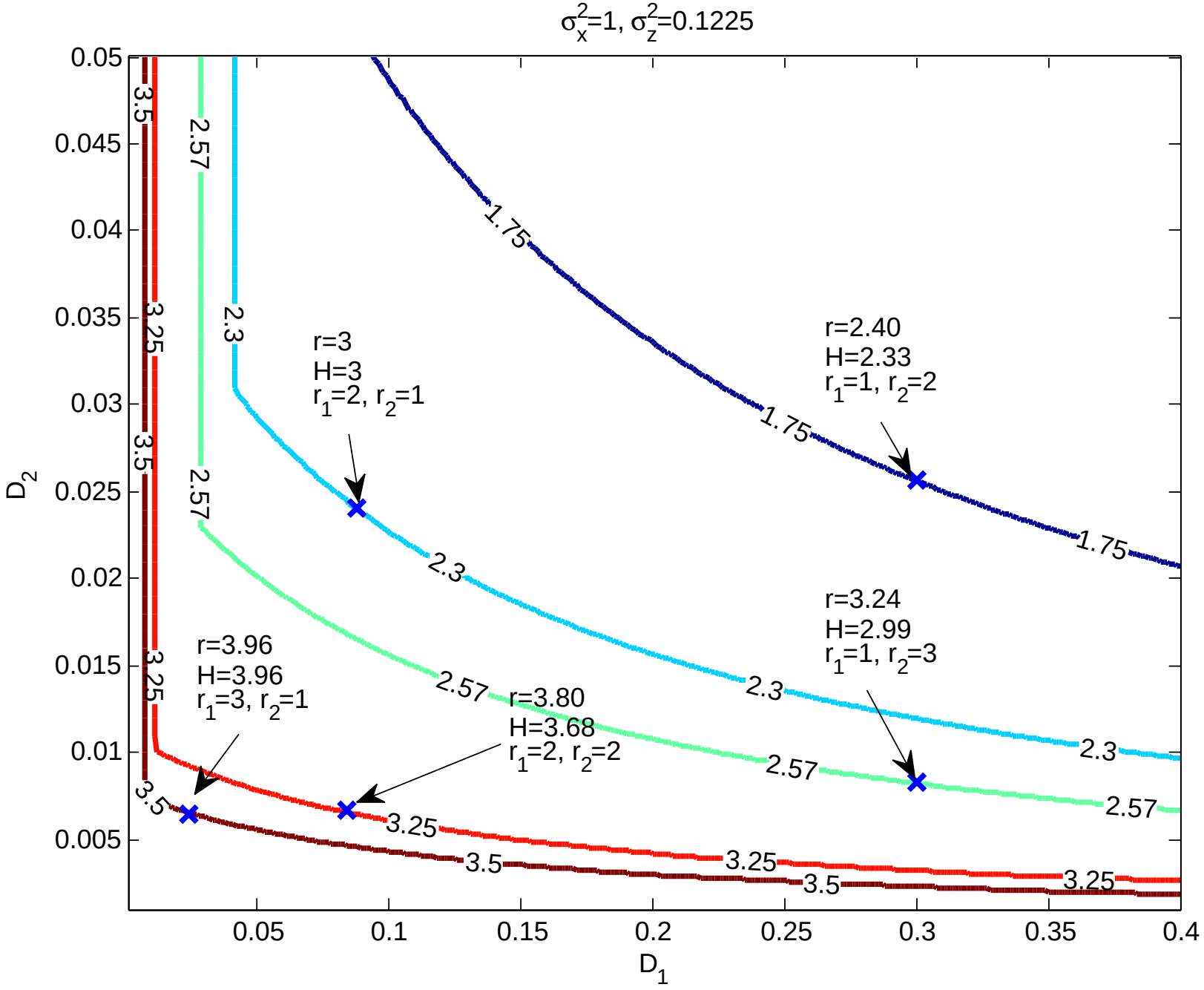
How might we do this? – ML & Practical Codes Part - I

- for this problem need both a *successive refinement* code and a *Wyner-Ziv* code. [18]
- correlated Gaussian $(\cdot)^2$ error
- start with $Y_1 = 0$
- SRTCQ [19, 20, 21]
- losslessly compress first refinement (for decoder 1)
- Take conditionally IID bits of refinement, use (now standard) LDPC syndrome based compression trick [22, 23]
- use a belief propagation iterative decoder using side information as observations to decode the syndrome [24]



- idea: iterate successive versions of these codes (different nodes taking turns)
- concatenate with a difficult inference model within the same BP decoder

How might we do this? – ML & Practical Codes Part - II



References

[1] S. Ramanan and J. M. Walsh, "Distributed Estimation of Channel Gains in Wireless Sensor Networks," in *Forty-Second Asilomar Conference on Signals, Systems, and Computers*, Oct. 2008. [Online]. Available: http://www.ece.drexel.edu/walsh/Ramanan_Asilomar_08.pdf

[2] S. Ramanan and J. M. Walsh, "Distributed Estimation of Channel Gains in Wireless Sensor Networks," *IEEE Trans. Signal Processing*, vol. 58, no. 6, pp. 3097–3107, June 2010. [Online]. Available: http://www.ece.drexel.edu/walsh/Ramanan_TSP_10.pdf

[3] M. Çetin, L. Chen, J. W. Fisher III, A. T. Ihler, R. L. Moses, M. J. Wainwright, and A. S. Willsky, "Distributed Fusion in Sensor Networks," *IEEE Signal Processing Mag.*, pp. 42–55, July 2006.

[4] A. T. Ihler, I. J. W. Fisher, R. L. Moses, and A. S. Willsky, "Nonparametric belief propagation for sensor network self-calibration," in *Proc. The International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Montreal, Quebec, May 2004.

[5] ——, "Nonparametric Belief Propagation for Self-Calibration in Sensor Networks," in *Information Processing in Sensor Networks (IPSN)*, July 2004.

[6] ——, "Nonparametric Belief Propagation for Sensor Network Self-Calibration," *IEEE J. Select. Areas Commun.*, vol. 23, april 2005.

[7] C. C. Moallemi and B. Van Roy, "Consensus propagation," *IEEE Trans. Inform. Theory*, vol. 52, no. 11, pp. 4753–4766, Nov. 2006.

[8] S. Amari and H. Nagaoka, *Methods of Information Geometry*. American Mathematical Society Translations of Mathematical Monographs, 2004, vol. 191.

[9] I. Csiszár and F. Matúš, "Information projections revisited," *IEEE Trans. Inform. Theory*, vol. 49, pp. 1474–1490, June 2003.

[10] R. L. Dobrushin and B. S. Tsybakov, "Information transmission with additional noise," *IEEE Transactions on Information Theory*, vol. IT-8, no. 5, pp. 293–304, September 1962.

[11] J. Chen, X. Zhang, T. Berger, and S. B. Wicker, "An upper bound on the sum-rate distortion function and its corresponding rate allocation schemes for the ceo problem," *IEEE Journal on Selected Areas in Communications*, vol. 22, no. 6, pp. 977–987, August 2004. [Online]. Available: http://dx.doi.org/10.1109/JSAC.2004.830888

[12] T. Berger, Z. Zhang, and H. Viswanathan, "The CEO Problem," *IEEE Transactions on Information Theory*, vol. 42, no. 3, pp. 887–902, May 1996.

[13] A. El Gamal and T. M. Cover, "Achievable rates for multiple descriptions," *IEEE Transactions on Information Theory*, vol. IT-28, no. 6, pp. 851–857, November 1982.

[14] A. Orlitsky, J. R. Roche, "Coding for Computing," *IEEE Trans. Inf. Theory*, vol. 47, no. 3, pp. 903–917, Mar. 2001. [Online]. Available: http://dx.doi.org/10.1109/18.915643

[15] Nan Ma, P. Ishwar, P. Gupta, "Information-theoretic bounds for multiround function computation in collocated networks," in *IEEE International Symposium on Information Theory (ISIT)*, 2009. [Online]. Available: http://dx.doi.org/10.1109/ISIT.2009.5205926

[16] Raymond W. Yeung, *Information Theory and Network Coding*. Springer, 2008.

[17] František Matúš, "Infinitely Many Information Inequalities," in *IEEE International Symposium on Information Theory (ISIT)*, June 2007, pp. 41–44.

[18] Sivagnanasundaram Ramanan and John MacLaren Walsh, "Practical Codes for Lossy Compression when Side Information May be Absent," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2011)*, May 2011, to appear.

[19] M. W. Marcellin and T. R. Fischer, "Trellis coded quantization of memoryless and gauss-markov sources," *IEEE Transactions on Communications*, vol. 38, no. 1, pp. 82–93, Jan. 1990.

[20] S. Sandeep Pradhan and K. Ramchandran, "Distributed source coding using syndromes (discus): Design and construction," *IEEE Transactions on Information Theory*, vol. 49, no. 3, pp. 626–643, Mar. 2003.

[21] H. Jafarkhani and V. Tarokh, "Design of successively refinable trellis-coded quantizers," *IEEE Transactions on Information Theory*, vol. 45, no. 5, pp. 1490–1497, Jul. 1999.

[22] A. D. Liveris, Z. Xiong, and C. N. Georghiades, "Compression of binary sources with side information at the decoder using ldpc codes," *IEEE Communications Letters*, vol. 6, no. 10, pp. 440–442, Oct. 2002.

[23] J. Garcia-Frias and Y. Zhao, "Compression of correlated binary sources using turbo codes," *IEEE Communications Letters*, vol. 5, no. 10, pp. 417–419, Oct. 2001.

[24] Y. Yang, S. Cheng, Z. Xiong, and W. Zhao, "Wyner-ziv coding based on tcq and ldpc codes," *IEEE Transactions on Communications*, vol. 57, no. 2, pp. 376–387, Feb. 2009.